PROTEIN INFORMATION RESOURCE FOR FUNCTIONAL GENOMICS AND PROTEOMICS

Cathy H. Wu
Department of Biochemistry & Molecular Biology, Georgetown University Medical Center, Box 571414, Washington, DC 20057-1414
wuc@georgetown.edu

The human genome project has revolutionized the practice of biology and the future potential of medicine. With the accelerated accumulation of high-throughput genomic and proteomic data, computational approaches are increasingly important for deriving scientific knowledge and hypotheses. There is a pressing need to develop advanced bioinformatics infrastructure for biological knowledge discovery. As an integrated public resource of protein informatics, the Protein Information Resource (PIR) provides many databases and analytical tools to support genomic and proteomic research and scientific discovery. The Protein Sequence Database (PSD) is the major annotated protein database in the public domain, containing about 280,000 sequences covering the entire taxonomic range. To provide high quality annotation and promote database interoperability, the PIR uses rule-based and classification-driven procedures based on controlled vocabulary and accepted ontologies, and includes evidence attribution to distinguish experimentally determined from predicted protein features. PIR-NREF, a non-redundant database containing almost 1,000,000 proteins from PIR-PSD, Swiss-Prot, TrEMBL, GenPept, RefSeq, and PDB, provides a timely and comprehensive sequence collection with source attribution for protein identification, ontology development of protein names, and detection of annotation errors. The iProClass database addresses the database interoperability issues arising from the voluminous, heterogeneous, and distributed data. It provides comprehensive family relationships and functional and structural features for about 800,000 proteins in PIR-PSD, Swiss-Prot, and TrEMBL, with rich links to over 50 databases of protein families, functions, pathways, protein-protein interactions, post-translational modifications, structures, genomes, ontologies, literature, and taxonomy. An integrated protein knowledgebase, connecting the underlying data warehouse and sequence analysis and data mining tools with graphical user interfaces, is being developed for large-scale gene expression and proteomic data analysis, functional categorization, and pathway identification. The PIR databases are implemented in an object-relational database system and accessible from our web site (http://pir.georgetown.edu) for exploration of proteins and their comparative analysis. It helps users to answer complex biological questions that may typically involve querying multiple sources and detect interesting relationships among protein sequences and groups. Such knowledge is fundamental to the understanding of protein evolution, structure, and function, and crucial to functional genomic and proteomic research.

The PIR is supported by the NIH grant P41 LM05798, iProClass is supported by the NSF grants DBI-9974855 and DBI-0138188, and the Protein Name Ontology project is supported by the NSF grant ITR-0205470.